Chapter 1. Introduction

The new technique of microarray is a high throughput technique to explore the expression profiles of a large amount of genes in genomics. This is an important tool to study the functionality of genes in the genomic scale in the era of post-genomics (Yang, Speed,2002). In order to measure the expression profiles of genes from microarrays, it is crucial to analyze the microarray images with high accuracy. Based on the these accurate measures, advanced analysis for selecting significant genes, clustering, classification, pathway and network reconstruction can be proceeded with solid foundations (Yang, Buckley, Dudoit and Speed,2002,Liu 2004, Li and Lu, 2005).

Microarray images record the fluorescence of hybridized probes and targets with dyes. The fluorescence is related to the expression level of RNAs in the samples. The most original information of microarray data is stored in microarray images. Therefore, it is important to evaluate the quality of microarray by inspecting the microarray image. Image processing measures the expression levels of mRNA on each spot of a microarray with four steps: image acquisition, spot location, computation of spot intensities and data reporting. Image analysis will involve the calibration of scanning efficiencies of dyes, the alignment and detection of spotting errors, the denoise of background errors, and the marking of dust, moving, hybridization, and other artifacts (Chen, Dougherty, and Bittner, 1997, Yang, Buckley, Dudoit, and Speed, 2002). Therefore, the processing of gridding, segmentation of foreground/background images, flags, normalization, smoothing, and statistics will affect the estimated ratios. We have investigated the process of gridding in our previous studies (Ho, Hwang, Lu, and Lee, 2005). This study is aimed to improve the segmentation of foreground/background images that will reduce the estimated ratios as a result. We will investigate two-channel microarray images in this study.

The segmentation of an image is the process of partitioning the image into different regions that have similar intensity levels and features within regions. The intensity levels and features will have distinction. In a microarray experiment, segmentation perform the classification of pixels as foreground or background. We will use the mixture models to fit two distributions of background and foreground intensities because

of the flexibility of mixture models with different parameters, including the location and scale parameters.

In order to evaluate the performance of segmentation results, we will investigate the estimated ratios for spike genes in microarrays with target ratios by mixture models and the software of GenePix 6.0. Spike genes are often used as the tools of quality control for microarrays. Spike genes are designed to have the artificial sequences that are different from the genes under studies. The sequence complements to spike genes are spotted on a microarray as probes in the process of printing. During the process of microarray hybridization, we will put in the other complement sequences as targets. The match pairs of probes and targets will be hybridized. For two-channel cDNA microarrays in this study, the ratios of Cy3 to Cy5 labeled targets are predefined for spike genes. These target ratios of spike genes will be the target values for estimated ratios by image processing of microarray images. Therefore, these target ratios of spike genes will be used to evaluate the performance of segmentation methods and image processing in our studies.

Chapter 2. Steps of Image Processing

We will review the main steps of image processing for microarray images and discuss the key factors that affect the performances.

2.1 Overview

The images of two-channel cDNA microarrays have two colors, like the red and green colors for Cy5 and Cy3 dyes respectively. These two dyes will be used to label control and experiment samples. The ratios of intensities in two colors for genes will be used to select differently expressed genes that have different expression levels among control and experiment samples. For instance, the ratio for one gene can be defined as the intensity of Cy5 divided by that of Cy3. If this ratio is much greater (or smaller) than 1, then this gene expresses more in the sample labeled with the Cy5 (or Cy3) dye and the color for the spot of this gene in the microarray image will be toward red (or green). These genes have differentially expressed levels and they are important for the distinction of control and experiment samples. On the other hand, if this ratio is close 1, then this gene expresses similar in both samples labeled with the Cy3 and Cy5 dye. Consequently, the color for the spot of this gene in the microarray image will be toward yellow. This kind of genes has non-differentially expressed levels. So, they are not useful for the distinction of control and experiment samples. Hence, it is very crucial to estimate the ratio accuarately in order to select the differentially expressed genes between control and experiment samples.

The commercial software of GenePix is widely used for microarray image processing. The latest version of GenePix Pro 6.0 will be used to evaluate the performance of current commercial software (see http://www.axon.com/GN_GenePixSoftware.html). GenePix 6.0 can read single or multiple TIFF images as inputs. By the built-in methods of image segmentation, GenePix can segment the foreground and background of every spot for one gene. Then, it report the features of every spot by calculating the ratio of mean, the ratio of median, the mean of ratio, the median of ratio, and other features. The built-in segmentation

methods in GenePix 6.0 include the methods of square, circular, and irregular boundaries.

2.2 Operation Steps in GenePix 6.0

In GenePix 6.0, the main standard operation procedure (SOP) contains 6 steps. Step 1 is loading image files. Step 2 is setting the numbers of block and gridding parameters for this image. Step 3 is to segment features for all blocks. Step 4 is to compute statistics of every spot. Step 5 is to normalize the results of features. Step 6 is reporting results. Some input arguments are needed before running the SOP of GenePix. Input arguments include the size of blocks, the number of blocks, the number of spots in a block, the size of a spot, and the methods of normalization as illustrated in Fig. 2.1, 2.2, and 2.3. The user needs to input and adjust these arguments manually in order to generat good results from GenePix. The effects and adjustments of these input arguments are discussed in Appendix I.

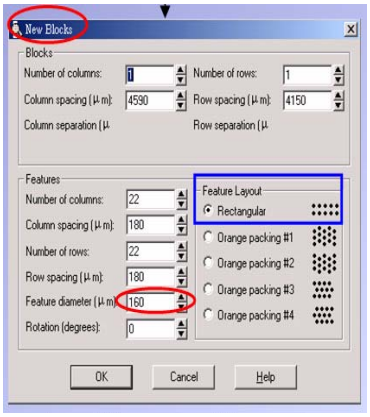


Fig. 2.1: Input arguments for image information.

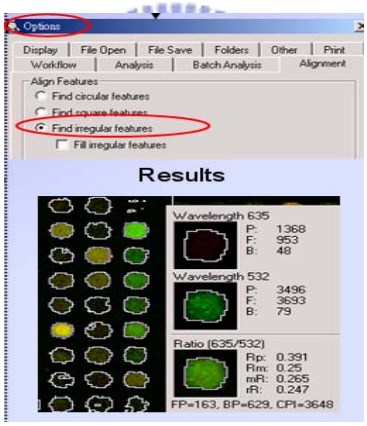


Fig. 2.2: Input arguments for selecting the method of segmentation.

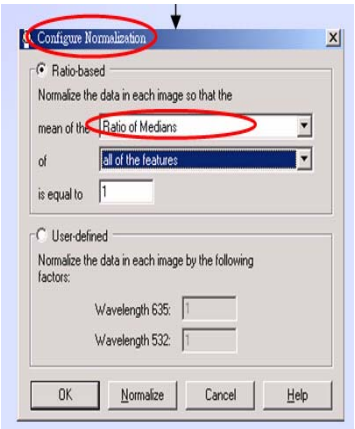


Fig. 2.3: Input arguments for selecting the method of normalization.

2.3 Performance of Segmentation in GenePix 6.0

Among the key steps in GenePix 6.0 shown in Figure 2.1, 2.2, and 2.3, the main challenge is the selection of a good method of segmentation. Typical segmentation results of one microarray image by GenePix 6.0 are displayed in Fig. 2.4. There are several spots that do not have accurate segmentation boundaries. Therefore, we are motivated to develop new methods to improve the performance of segmentation for microarray images.

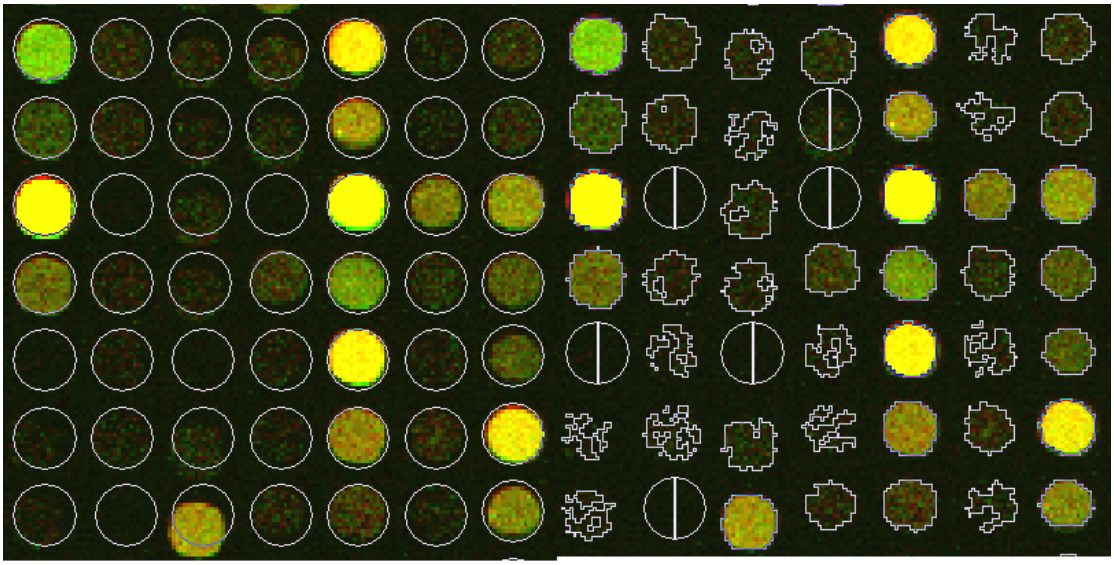


Fig. 2.4: Typical segmentation results of one microarray image by GenePix 6.0 are displayed.

Chapter 3. Image Segmentation by Mixture Models

The histogram of pixel intensities in one microarray image is illustrated in Figure 3.1, which reveals that there are two major distributions for foregrounds and backgrounds. Hence, we can model the distribution of pixel intensities by mixture models that have different location and scale parameters for foregrounds and backgrounds. Then, we can estimate the cut point of these two distributions to segment image pixels to foregrounds and backgrounds. For model simplicity, we will consider the normal mixture model (NMM) in this study. Mixture models of other distributions are possible in future studies.

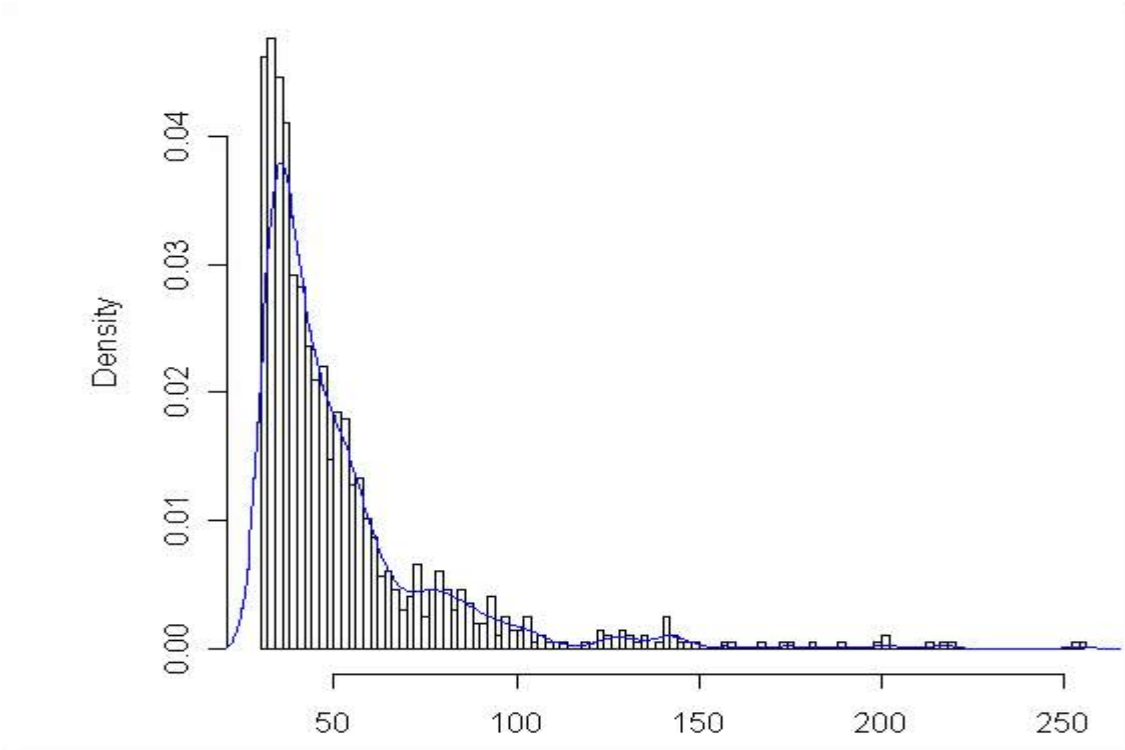


Fig. 3.1: The histogram of pixel intensities in one microarray image is shown.

3.1 Flowchart

In the new method of segmentation by NMM proposed in this study, the main operation procedure contains seven steps. Step1 is loading two image files with Cy3 and Cy5 separately. Step 2 is combining two intensities of Cy3 and Cy5 images at one pixel into the average intensity. Step 3 is to segment features using the image values of

the combined image by NMM. Step 4 is finding out the boundary for each spot. Step 5 is performing smoothing for each spot. Step 6 is finding out the normalization factor. Step 7 is reporting results. In order to make comparisons with GenePix 6.0, we will use the same input arguments and coordination in the operation process as GenePix 6.0 whenever feasible. Figure 3.2 displays the flowchart of the new method by NMM. In the following sections, we will explain the detail elements of flowchart.

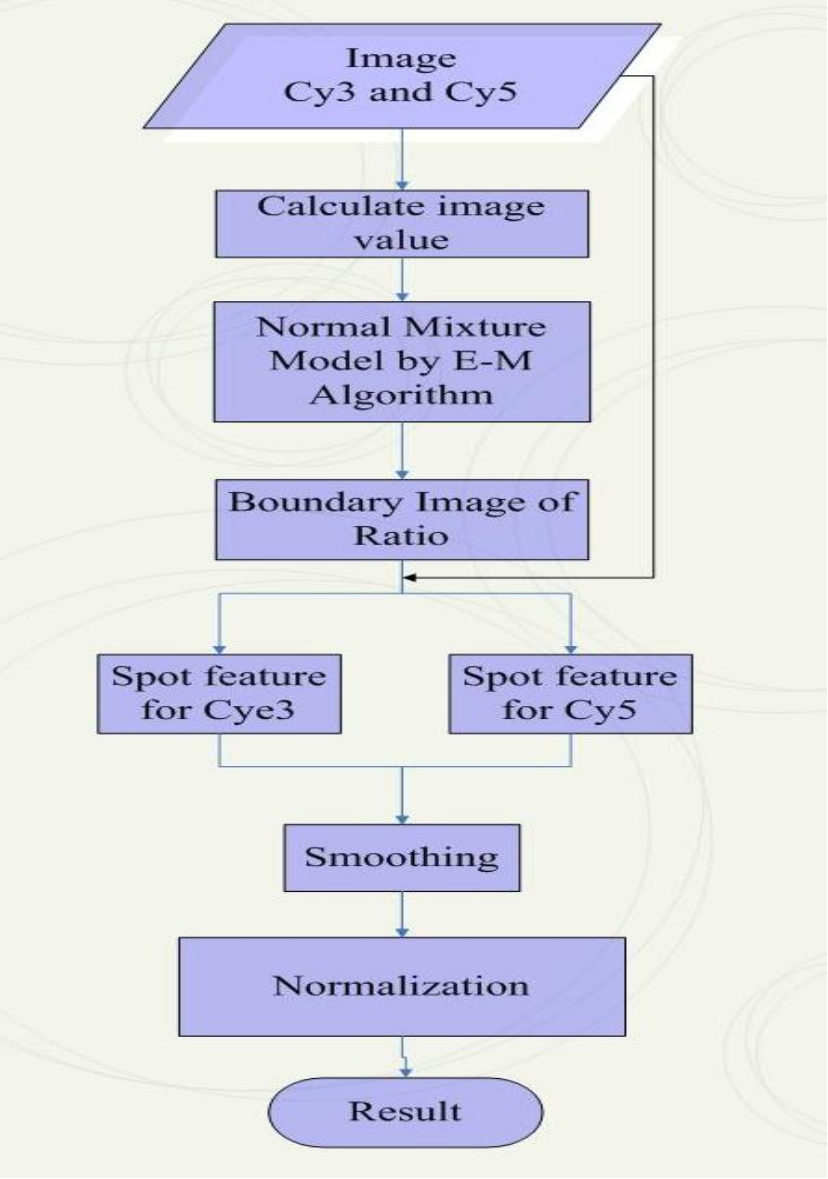


Fig. 3.2: Flowchart of the proposed method segmentation of microarray images by NMM is illustrated.

3.2 Combined Images

Most of microarray images with TIFF format use 2 bytes to store the intensity of a pixel. Therefore, the intensity of each pixel is between 0 and 65535. It can be transformed to 0 and 255 for the purpose of display in a computer monitor. In order to combine two images of Cy3 and Cy5, the average of two image values in Cy3 and Cy5 images will be used as the image value of the combined image for each pixel. Using the intensities of combined image, we can segment the foreground and background pixels in every spot by NMM.

The advantage of using the combined image value for segmentation is that we can have the same segmentation of foreground and background pixels for both Cy3 and Cy5 images. Then, we can calculate the estimated ratios by different approaches accordingly without any problem, including the ratio of mean, the ratio of median, the mean of ratio, the median of ratio, and other features. If the segmentation of Cy3 and Cy5 images are performed separately, then it is very likely the segmentation boundaries of every spot are different in Cy3 and Cy5 images. This may cause problems in calculating statistics in the level of pixels.

3.3 EM Algorithm for a Normal Mixture Model (NMM)

We suppose the distribution of foreground intensities follows a normal distribution of $f_1(\mu_1, \sigma_1^2)$ with mean μ_1 and variance σ_1^2 and the distribution of background intensities follows another normal distribution of $f_2(\mu_2, \sigma_2^2)$ with mean μ_2 and variance σ_2^2. Therefore the distribution of one pixel x_j in a spot can be model as a mixture of two normal distributions.

$$f(x_j;\phi) = \pi_1 f_1(x_j;\mu_1,\sigma_1^2) + \pi_2 f_2(x_j;\mu_2,\sigma_2^2), j = 1,...,n, \tag{3.3.1}$$

where

$$f_1(x_j;\mu_1,\sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-(x_j-\mu_1)^2}{2\sigma_1^2}\right),$$

$$f_2(x_j;\mu_2,\sigma_2^2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{-(x_j-\mu_2)^2}{2\sigma_2^2}\right),$$

$$\phi = \{\pi_1, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2\},$$

and $\pi_i, i = 1,2,$ is the mixing proportion (or prior probability) for the foreground and background subject to the constraints of

$$0 \le \pi_i \le 1 \quad \text{and} \quad \pi_1 + \pi_2 = 1.$$

As the foreground intensities conatin the signals and noises, the mean of foreground intensities is typically greater than that of background intensities. Hence, we can consider the identifiable condition that $\mu_1 \ge \mu_2$ (McLachlan and Peel, 2000).

The log-likelihood of observed data in the model of two mixtures becomes

$$\log(L(\phi \mid x)) = \sum_{j=1}^{n} \log(\sum_{i=1}^{2} \pi_i f_i(x_j; \mu_i, \sigma_i^2)). \tag{3.3.2}$$

The maximum likelihood estimate can be estimated by solving the partial differential equations of $\partial \log(L(\phi \mid x)/\partial \phi = 0$. However, this approach will encounter numerical difficulties. The iterative method of EM algorithm can be applied to estimate parameters by a simpler approach numerically. The algorithm has two steps, E (Expectation) and M (Maximization) steps. By introducing the unobserved indicator variables of Z_{ij}, the log likelihood function of the complete data turns out to be

$$\log(L_c(\phi \mid x)) = \sum_{j=1}^{n} \sum_{i=1}^{2} Z_{ij} (\log \pi_i + \log f_i(x_j; \mu_i, \sigma_i^2)). \tag{3.3.3}$$

where $Z_{ij} = \begin{cases} 1, & \text{when } x_j \text{ is from the } i\text{th distribution;} \\ 0, & \text{otherwise.} \end{cases}$

E-Step:

In the E-step, the mixing parameter π_i can be thought as the prior probability of each mixture component. By the Bayes rule, the posterior probability that x_j belongs to the *i*th distribution of the mixture given the current estimates of parameters at *k*th iteration becomes

$$P(i \mid x_j; \phi^{(k)}) = \frac{\Pr(x_j, i; \phi^{(k)})}{\Pr(x_j; \phi^{(k)})} = \frac{\pi_i f_i(x_j; \mu_i^{(k)}, \sigma_i^{2(k)})}{\sum_{i=1}^{2} \pi_i f_i(x_j; \mu_i^{(k)}, \sigma_i^{2(k)})} = \tau_{ij}^{(k)} \tag{3.3.4}$$

Also, the conditional expectation of $\log L_c(\phi \mid x)$ is as follows:

$$Q(\phi;\phi^{(k)}) = E_{\phi^{(k)}}(\log L_c(\phi \mid x))$$

$$= E_{\phi^{(k)}}\left[\sum_{j=1}^{n}\sum_{i=1}^{2} Z_{ij}(\log \pi_i + \log f_i(x_j;\mu_i,\sigma_i^2))\right]$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{2} E_{\phi^{(k)}}[Z_{ij} \mid x](\log \pi_i + \log f_i(x_j;\mu_i,\sigma_i^2))$$

$$= \sum_{j=1}^{n}\sum_{i=1}^{2} \tau_{ij}^{(k)}(\log \pi_i + \log f_i(x_j;\mu_i,\sigma_i^2)) \qquad (3.3.5)$$

M-Step:

In the M-step, one will maximize $Q(\phi;\phi^{(k)})$ in (3.3.5) subject to the constraint that $\sum_{i=1}^{2}\pi_i = 1$. The Lagrange method can incorporate Q function with the constraint as follows:

$$Q_\lambda(\phi;\phi^{(k)}) = Q(\phi;\phi^{(k)}) + \lambda(\sum_{i=1}^{2}\pi_i - 1) \qquad (3.3.6)$$

where λ is a Lagrange multiplier. By maximizing the above Lagrange functional, the estimates of the mixture proportions and other parameters turn out to be

$$\pi_i^{(k+1)} = \frac{1}{n}\sum_{j=1}^{n}\tau_{ij}^{(k)}, \qquad (3.3.7)$$

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(k)}x_j}{\sum_{j=1}^{n}\tau_{ij}^{(k)}}, \qquad (3.3.8)$$

$$\sigma_i^{2(k+1)} = \frac{\sum_{j=1}^{n}\tau_{ij}^{(k)}(x_j - \mu_i^{(k+1)})^2}{\sum_{j=1}^{n}\tau_{ij}^{(k)}}. \qquad (3.3.9)$$

It is well known that the EM algorithm is a simple and iterative algorithm with row operation and linear complexity. Furthermore, the log-likelihood is non-decreasing after every iteration of the EM-algorithm, which leads to montoical convergence under regular conditions (Dempster, Laird, and Rubin, 1977, Wu, 1983, McLachlan and Peel, 2000).

The EM algorithm of two normal mixtures in this study is listed as follows.

Step 1: Input initial parameters: $k = 0$,

$\pi_1 = 0.25, \pi_2 = 1 - \pi_1,$

$\mu_1 = Q_1, \mu_2 = Q_3,$ (where Q_1=the 25th quintile, Q_3=the 75th quintile),

$\sigma_1 = \sigma_2 = 0.5$ *(sample standard deviation).

Step 2: Calculate $\tau_{ij}^{(k)} = \dfrac{\pi_i f_i(x_j; \mu_i^{(k)}, \sigma_i^{2(k)})}{\sum_{m=1}^{2} \pi_m f_m(x_j; \mu_m^{(k)}, \sigma_m^{2(k)})}$.

Step 3: Calculate new estimates of $\pi_i^{(k+1)} = \dfrac{1}{n}\sum_{j=1}^{n} \tau_{ij}^{(k)}$, $\mu_i^{(k+1)} = \dfrac{\sum_{j=1}^{n} \tau_{ij}^{(k)} x_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}$, and

$$\sigma_i^{2(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} (x_j - \mu_i^{(k+1)})^2}{\sum_{j=1}^{n} \tau_{ij}^{(k)}}.$$

Step 4: If $\log(L_c(\phi^{(k+1)} \mid x)) - \log(L_c(\phi^{(k)} \mid x)) < tol$ and stop, otherwise k <- $k+1$ and go to Step 2.

When we obtain the maximum likelihood estimates of parameters by above EM algorithm, we can use the NMM to segment image pixels by finding the cut point between the distributions of foreground and background intensities. If a pixel has intensity greater than the cut point, then it will be allocated to the foreground. Otherwise, that pixel is allocated to the background.

We will consider the cut point x^* such that

$$\pi_1 f_1(x^*; \mu_1, \sigma_1^2) = \pi_2 f_2(x^*; \mu_2, \sigma_2^2). \qquad (3.3.10)$$

Thus, the cut point becomes

$$x^* = \frac{(\sigma_1^2\mu_2 - \sigma_2^2\mu_1) \pm \sqrt{(\sigma_2^2\mu_1 - \sigma_1^2\mu_2)^2 - (\sigma_1^2 - \sigma_2^2)\left[\left(\sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2\right) - 2\sigma_1^2\sigma_2^2 \log(\frac{\pi_2\sigma_1}{\pi_1\sigma_2})\right]}}{(\sigma_1^2 - \sigma_2^2)}, \qquad (3.3.11)$$

when $\sigma_1^2 \neq \sigma_2^2$ and the square root in the numerator exists. When $\sigma_1^2 = \sigma_2^2$ and $\mu_1 \neq \mu_2$, the cut point becomes

$$x^* = \frac{\sigma_1^2}{(\mu_1 - \mu_2)}(\log\frac{\pi_2}{\pi_1} + \frac{\mu_1^2 - \mu_2^2}{2\sigma_1^2}). \qquad (3.3.12)$$

define the cut point equal the median of intensities in every spot.

3.4 Gaussian Smoothing

The purpose of Gaussian smoothing is to denoise image intensities using spatial information and improve the accuracy for estimated features after segmentation. The Gaussian smoothing operator is a convolution operator that is used to smooth images and reduce noises. This is a process that data points are averaged with their neighbors according to the weights in the Gaussian kernel.

There are two kinds of parameters in Gaussian smoothing. They are the size of mask window of neighboring pixels and the scales of the standard deviances in the Gaussian kernel. The Gaussian kernel is a Gaussian distribution in the mask window. For two-dimensional images in this study, we will consider the Gaussian distribution with the same scales of the standard deviances in both horizontal and vertical directions because there are no distinguishable variations in these two scales for the generation process of microarray images. The details are discussed in Appendix II.

3.5 Normalization Factors

The purpose of normalization is to remove the systematic effects, like the effects of dyes, arrays, blocks, print-tips, and so forth. We will consider the normalization of dye effects in this study for simplicity. We will use the geometric means of block medians in Cy3 and Cy5 images (Cy3-BlockMedian i and Cy5-BlockMedian i) to avoid the effects of extreme values. Hence, we will use the following normalization factor for case studies:

$$Norm.Factor = \frac{\prod_{i=1}^{B}(Cy5\_BlockMedian_i)^{1/B}}{\prod_{i=1}^{B}(Cy3\_BlockMedian_i)^{1/B}}, \qquad (3.5.1)$$

where B is the number of blocks in an image.

Chapter 4. Empirical Studies

Microarray images in this study are provided by Dr. Yun-Shien Lee in C. G. M. H. Each image has 32 blocks with 22 columns and 22 rows. An example of microarray image is shown in Figure 4.1. Eight spike genes are spotted in each block. One block is shown by enlarging and the spike genes are segmented by the rectangle in Figure 4.2.

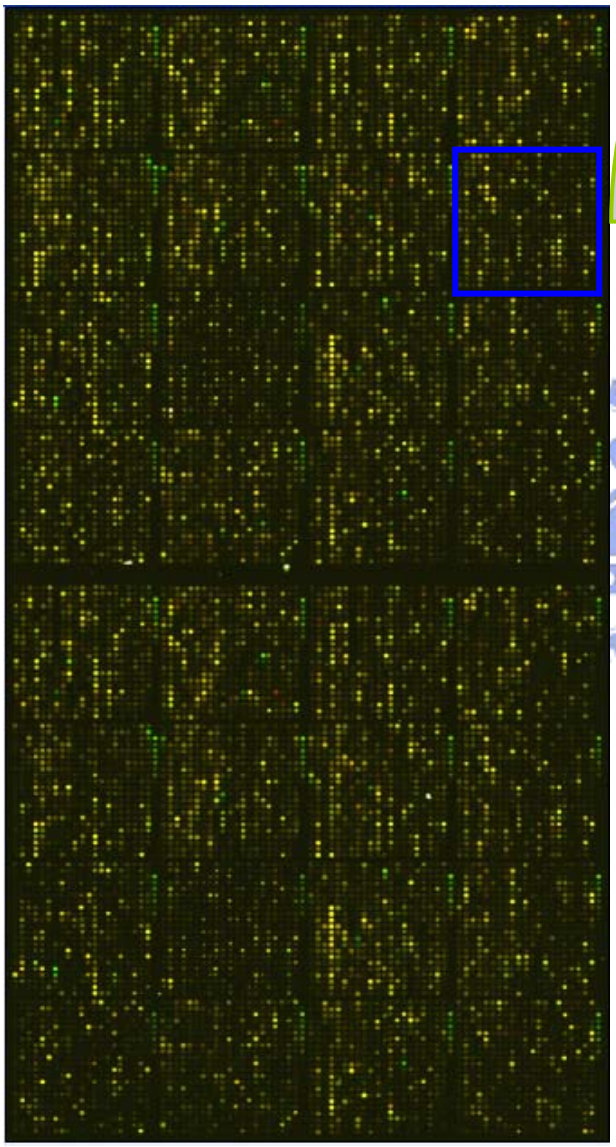


Figure 4.1: An example of microarray Image is shown.

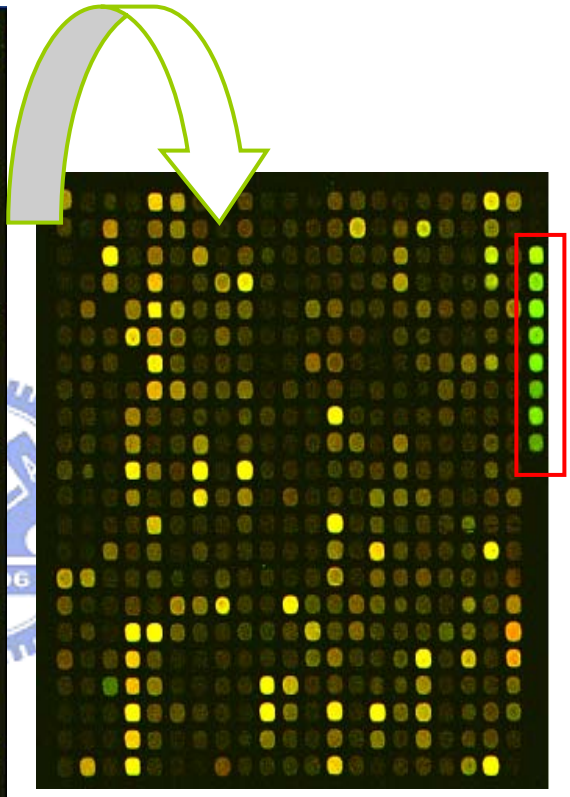


Figure 4.2: One block in Figure 4.1 enlarged and the spike genes are highlighted in the rectangle.

4.1 Data and Statistics

The target ratios of eight spike genes are reported in Table 4.1. These provide the golden standard for evaluating the performance of the segmentation method by NMM proposed in this study. This also serves as the common platform for the comparisons of the proposed method with those in GenePix 6.0.

Table4.1: The target ratios of eight spike genes are listed.

Spike Gene	1	2	3	4	5	6	7	8
Target Content (635:532)	50:500	10:100	50:250	20:100	200:500	40:100	200:200	20:20
Target Ratio (635:532)	1:10	1:10	1:5	1:5	1:2.5	1:2.5	1:1	1:1

We will calculate two estimated ratios by NMM. Then, we will evaluate the sum of square of relative errors (SSREs) for GenePix6.0 and NMM. The formulas for two estimated ratios are defined in (4.1.1) and (4.1.2). The sum of square of relative errors (SSREs) between target ratios (TR) and the averages of estimated ratios (AER) of seven spike genes with 32 replications in a image is used to evaluate and compare the performance of segmentation methods. The SSRE is used to evaluate and compare the accuracy of segmentation as defined in (4.1.3). If SSREs are smaller, the estimated ratios are closer to target ratio. The detail formulas are given below:

Let $I_{P,\lambda}$: The pixel intensity of single spot for laser λ,

$I_{B,\lambda}$: The background intensity of single spot for laser λ,

λ: Cy5 or Cy3,

$\langle S_k \rangle_{med}$: the median of k observed data in S,

n: the number of spots in foreground,

m: the number of spots in background.

Then, we will have the followings:

$$\text{The ratio of median: } \frac{\left\langle \left(I_{P,Cy5} \right)_n \right\rangle_{med} - \left\langle \left(I_{B,Cy5} \right)_m \right\rangle_{med}}{\left\langle \left(I_{P,Cy3} \right)_n \right\rangle_{med} - \left\langle \left(I_{B,Cy3} \right)_m \right\rangle_{med}} \qquad (4.1.1)$$

II. The ratio of mean: $$\frac{\sum_{i=1}^{n}\left(I_{P,Cy5}-\left\langle\left(I_{B,Cy5}\right)_m\right\rangle_{med}\right)_i}{\sum_{i=1}^{n}\left(I_{P,Cy3}-\left\langle\left(I_{B,Cy3}\right)_m\right\rangle_{med}\right)_i} \quad (4.1.2)$$

$$\text{SSRE} = \sum_{i=1}^{8}\left(\frac{i^{th}AER}{i^{th}TR}-1\right)^2, \quad (4.1.3)$$

where AER = Average Estimated Ratio and TR = Target Ratio.

4.2 Segmentation Results by GenePix 6.0 and NMM

In the segmentation results by GenePix 6.0, this study will use the feature diameter of 160 μm, normalization factor by the global ratio of median, background subtraction by global mean of all feature background medians, and irregular segmentation method without filled spots. Table 4.2 reports the results of SSRE in dye swap or one single microarrays under test by GenePix 6.0. The chip name of 62N62T means that the normal (and tumor) sample is labeled by the Cy5 (and Cy3) dye. We can change the configurations, factors, and statistics in image analysis software to minimize the SSRE. The software of GenePix 6.0 is used to perform this evaluation in this study. The results illustrate the feasibility of this proposed procedure.

Table 4.2: The results of SSREs for test images by GenePix 6.0 are listed.

Image		**62N62T**		**62T62N**	
Spike#	T. Ratio	Ratio of Median (635/532)	Ratio of Means (635/532)	Ratio of Median (635/532)	Ratio of Means (635/532)
Spike1	0.1	0.11	0.11	0.41	0.43
Spike2	0.1	0.05	0.06	0.17	0.22
Spike3	0.2	0.08	0.09	0.31	0.34
Spike4	0.2	0.10	0.12	0.39	0.44
Spike5	0.25	0.18	0.19	0.65	0.66
Spike6	0.25	0.20	0.22	0.78	0.83
Spike7	1	0.68	0.65	2.43	2.36
Spike8	1	0.48	0.44	3.40	2.04
SSRE		1.3287	**1.1613**	26.0087	**25.2790**

Image		63T63N		63N63T	
Spike#	T. Ratio	Ratio of Median (635/532)	Ratio of Means (635/532)	Ratio of Median (635/532)	Ratio of Means (635/532)
Spike1	0.1	0.14	0.15	0.32	0.35
Spike2	0.1	0.05	0.08	0.14	0.18
Spike3	0.2	0.11	0.13	0.25	0.28
Spike4	0.2	0.13	0.16	0.32	0.36
Spike5	0.25	0.23	0.24	0.54	0.55
Spike6	0.25	0.28	0.30	0.66	0.71
Spike7	1	0.94	0.87	2.10	2.01
Spike8	1	1.12	0.72	3.03	1.67
SSRE		0.7738	**0.5894**	14.86456	**14.09379**

Image		54T54N	
Spike#	T. Ratio	Ratio of Median (635/532)	Ratio of Means (635/532)
Spike1	0.1	0.31	0.32
Spike2	0.1	0.12	0.16
Spike3	0.2	0.22	0.25
Spike4	0.2	0.27	0.32
Spike5	0.25	0.49	0.51
Spike6	0.25	0.57	0.62
Spike7	1	1.94	1.87
Spike8	1	2.45	1.67
SSRE		10.1228	**10.0081**

In the segmentation results by NMM, we define coordinates and diameters of every spot the same as GenePix 6.0. Both in the segmentation results of GenePix 6.0 and NMM, the backgrounds are selected by the same method. Namely, the backgrounds of every spot are selected from the backgrounds in the region between the inner circle of one time of feature diameter that cover the foregrounds and the outer circle of three times of feature diameter. Table 4.3 reports the results of SSRE in dye swap or one single microarrays under test by NMM.

Table 4.3: The results of SSREs for test images by NMM are listed.

Image		62N62T		62T62N	
Spike#	T. Ratio	Ratio of Median (635/532)	Ratio of Means (635/532)	Ratio of Median(635/532)	Ratio of Means (635/532)
Spike1	0.1	0.12	0.13	0.4	0.41
Spike2	0.1	0.06	0.07	0.18	0.2
Spike3	0.2	0.09	0.1	0.3	0.32
Spike4	0.2	0.12	0.13	0.38	0.41
Spike5	0.25	0.2	0.21	0.63	0.63
Spike6	0.25	0.23	0.24	0.74	0.77
Spike7	1	0.75	0.73	2.3	2.23
Spike8	1	0.54	0.49	2.6	1.97
SSRE		0.9954	**0.9149**	**21.1963**	21.5488

Image		63T63N		63N63T	
Spike#	T. Ratio	Ratio of Median (635/532)	Ratio of Means (635/532)	Ratio of Median (635/532)	Ratio of Means (635/532)
Spike1	0.1	0.14	0.15	0.29	0.31
Spike2	0.1	0.06	0.08	0.13	0.16
Spike3	0.2	0.11	0.13	0.22	0.25
Spike4	0.2	0.13	0.16	0.28	0.31
Spike5	0.25	0.23	0.25	0.46	0.48
Spike6	0.25	0.28	0.31	0.56	0.59
Spike7	1	0.94	0.89	1.81	1.74
Spike8	1	1.02	0.77	1.89	1.61
SSRE		0.6698	**0.5751**	**7.5712**	8.8053

Image		54T54N	
Spike#	T. Ratio	Ratio of Median (635/532)	Ratio of Means (635/532)
Spike1	0.1	0.28	0.29
Spike2	0.1	0.11	0.14
Spike3	0.2	0.20	0.22
Spike4	0.2	0.24	0.28
Spike5	0.25	0.45	0.45
Spike6	0.25	0.50	0.53

Spike7	1	1.72	1.66
Spike8	1	1.72	1.48
SSRE		**5.9668**	6.5004

4.3 Comparisons of Results by GenePix 6.0 and NMM

We compare the SSREs of GenePix 6.0 and NMM for the same test image. The comparison results of SSREs of GenePix 6.0 and NMM for test images are reported in tables of Appendix III. Then, we will select the minimum of SSRE by GenePix 6.0 to decide which statistics shall be used for one test image. Once we decide the selected statistics for one test image, the AERs by GenePix 6.0 and NMM for spike genes are plotted in Figure 4.3-4.7 for 8 spike genes. From these figures, the results of AERs for spike genes by NMM are typically more close to target ratios than those by GenePix 6.0 are.

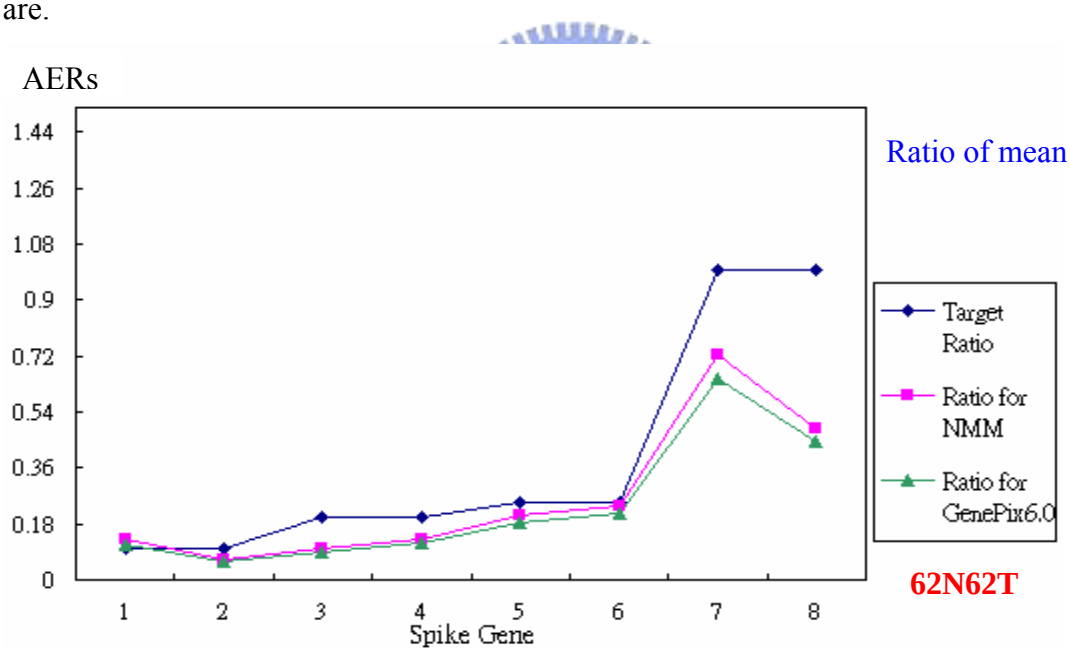


Figure4.3: The AERs by GenePix 6.0 and NMM for spike genes are plotted.

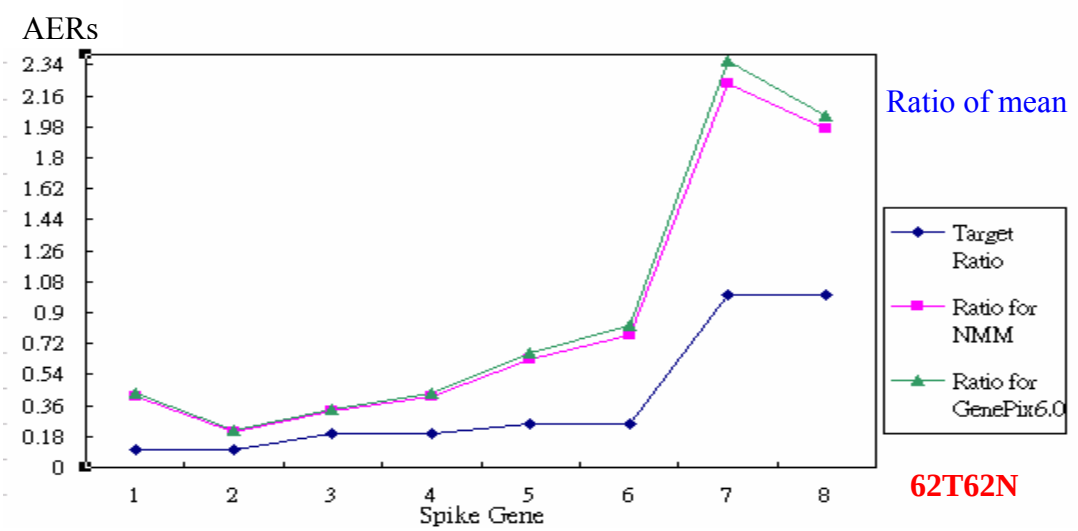


Figure4.4: The AERs by GenePix 6.0 and NMM for spike genes are plotted.

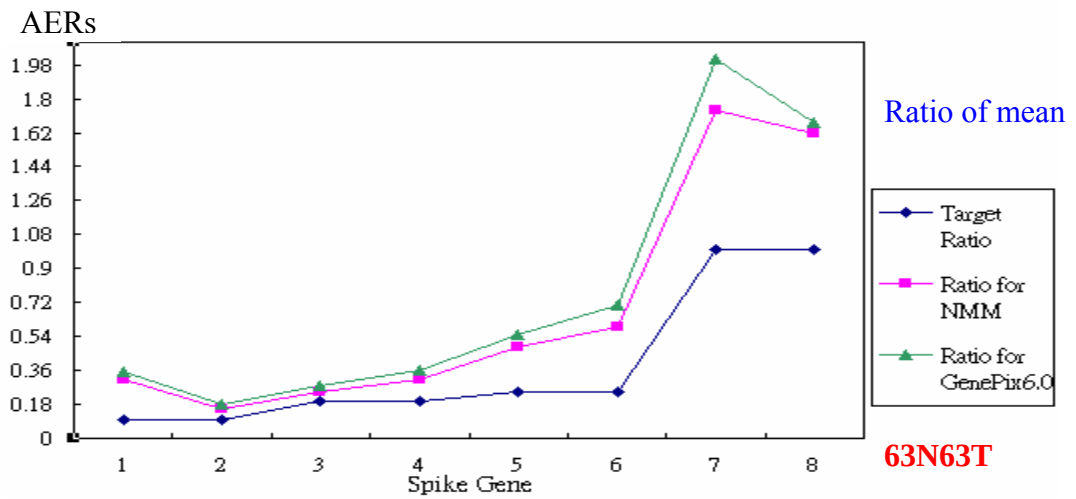


Figure4.5: The AERs by GenePix 6.0 and NMM for spike genes are plotted.

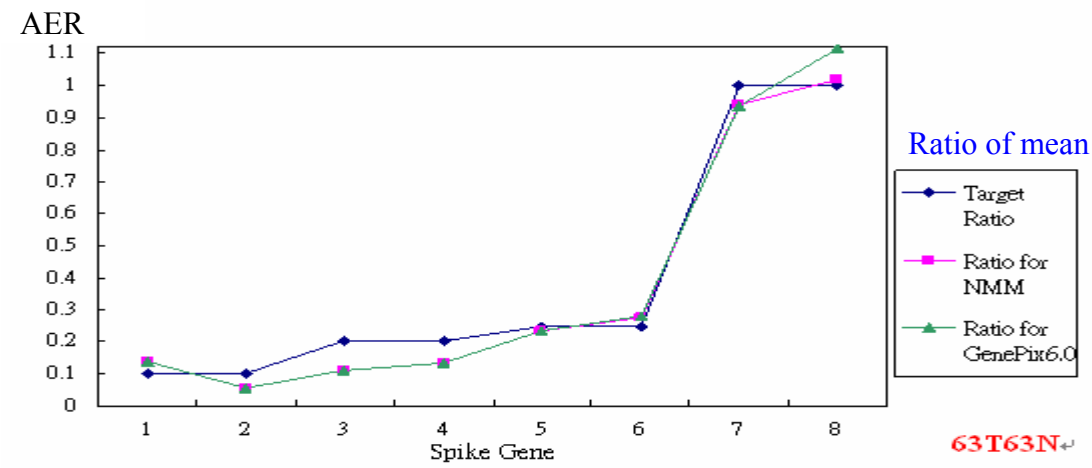


Figure4.6: The AERs by GenePix 6.0 and NMM for spike genes are plotted.

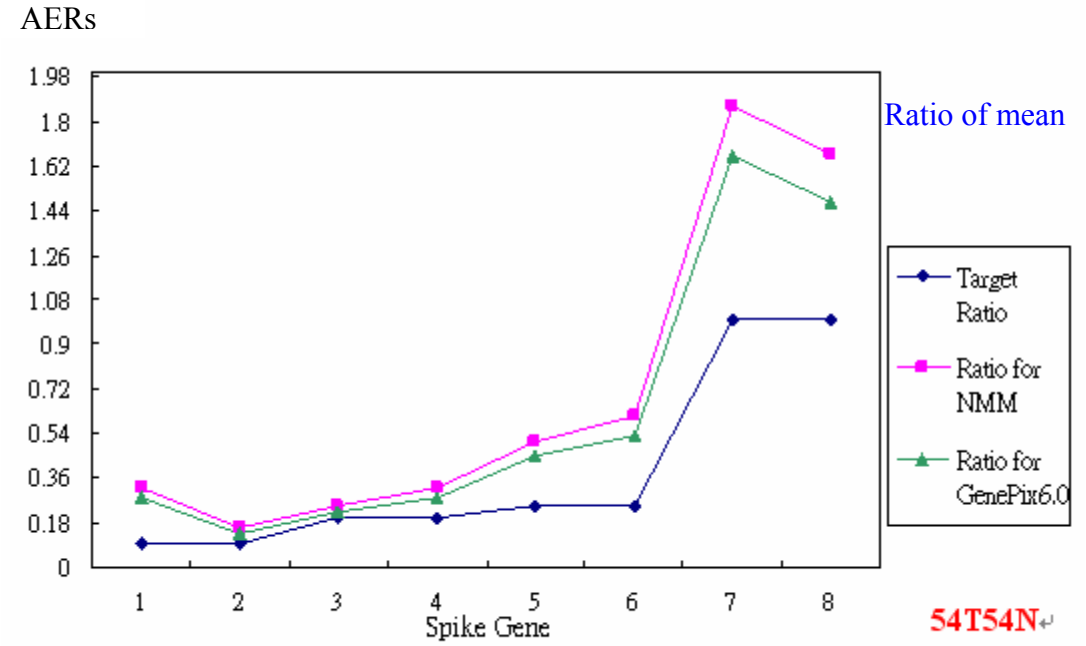


Figure4.7: The AERs by GenePix 6.0 and NMM for spike genes are plotted.

In Figure 4.6, we found the AER. of spike gene 8 in NMM is more close to target ratios than that of GenePix 6.0 is. Four typical spots of spike gene 8 in this microarray are enlarged in Figure 4.8. The middle row displays the original image. The upper row displays the segmentation results by GenePix 6.0 and these segmentations contain dark pixels that shall belong to backgrounds. Hence, the AERs by GenePix 6.0 are not close to the target ratios. The bottom row displays the segmentation results by NMM, which exclude dark pixels that shall belong to backgrounds. Hence, the AERs of NMM are close to target ratios.

The numerical comparisons of SSREs by GenePix 6.0 and NMM are reported in Table 4.4. The NMM method can reduce the SSEs between 2-37%.

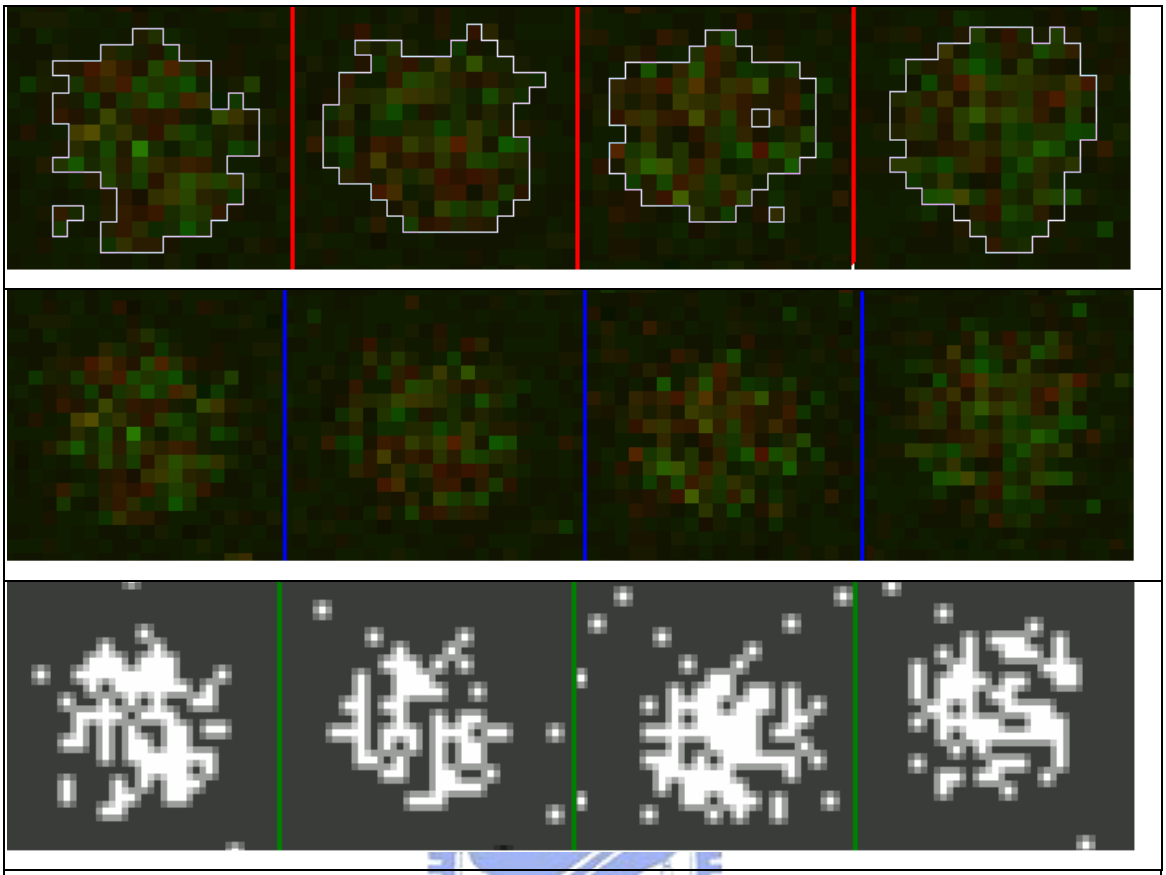


Figure4.8: Four typical spots of spike gene 8 in this microarray are enlarged.

Table 4.4: The comparisons of SSREs in test images by GenePix6.0 and NMM are listed.

SSRE	GenePix 6.0	NMM	Reduction
63T63N	0.5894	0.5751	**2.43%**
63N63T	14.0938	8.8053	**37.50%**
62T62N	25.2790	21.5488	**14.76%**
62N62T	1.1613	0.9149	**21.22%**
54T54N	10.0081	6.5004	**35.05%**

Chapter 5. Conclusion and Discussion

We have proposed a new segmentation method by the NMM to further improve the segmentation of microarray images. The mixture models are flexible for modeling the distributions of foreground and background intensities with different location and scale parameters. Empiric studies have confirmed the reduction of SSREs in comparison to the commercial software of GenePix 6.0.

The statistics of the mean of ratio and the median of ratio may have negative values when foreground intensities are smaller than background intensities. Further investigation of this kind of statistics is necessary to explore better statistics.

We use the cut point of foreground and background distributions to segment microarray images. We can also try to use the posterior probability in (3.3.4) to segment microarray images in future studies. Other mixtures besides normal mixtures are also possible (McLachlan and Peel 2000). Mixture number can be selected by the criteria of model selection, like the Bayesian information criterion and others (Schwrtz.1978). However, the decision of foreground for more than two mixtures will need other criterion.

Spatial information can be integrated to improve segmentation. For instance, we can select connected regions for foreground to remove isolated pixels. Other methods for segmentation of images can be further investigated (Chen, Lu, and Lin 2001, Chen, Lu, and Huang 2002, Wu, and Lu, 2004). We also plan to study the possibility to apply the segmentation of NMM in one-channel microarray images and other types of images..

References

1. Aitkin, M., Rubin DB. Estimation and hypothesis test in finite mixture models. JRSS Series B. Vol. 47 No. 1 67-75. 1985.
2. Amaratunga, Dhammika and Javier, Cabrera.Exploration and Analysis of DNA Microarray and Protein Array Data.. Hoboken, NJ: Wiley-Interscience-John Wiley and Sons, Inc. 2004.
3. Buhler, J., Ideker, T. and Haynor, D. Improved techniques for finding spots on cDNA microarrays, University of Washington, 2000.
4. Celeux, G. and Govaert, G.. A classification EM algorithm for clustering and two stochastic versions. Computational Statistics and Data Analysis, 14(3) :315–332. 1999.
5. Chen, Y., Dougherty, E. R., and Bittner, M. L. Ratio based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics 2, 364-374. 1997.
6. Chen, C.-M., Lu, H. H.-S., and Lin, Y.-C. An Early Vision Based Snake Model for Ultrasound Image Segmentation. Ultrasound in Medicine and Biology, 26, 2, 273-285. 2000.
7. Chen, C.-M., and Lu, H. H.-S. An Adaptive Snake Model for Ultrasound Image Segmentation: Modified Trimmed Mean Filter, Ramp Integration and Adaptive Weighting Parameters. Ultrasonic Imaging, 22,214-236. 2001.
8. Chen, C.-M., Lu, H. H.-S., and Hsiao, A.-T. A Dual Snake Model of High Penetrability for Ultrasound Image Boundary Extraction. Ultrasound in Medicine and Biology, 27, 12, 1651-1665. 2001.
9. Chen, C.-M., Lu, H. H.-S., and Huang, Y.-S. Cell-Based Dual Snake Model: A New Approach to Extracting Highly Winding Boundaries in The Ultrasound Images. Ultrasound in Medicine and Biology, 28, 8, 1061-1073. 2002.
10. Chen, C.-M., Lu, H. H.-S., and Chen, Y.-L. A Discrete Region Competition Approach Incorporating Weak Edge Enhancement for Ultrasound Image Segmentation. Pattern Recognition Letters, 24, 693-704. 2003.
11. F.H.C. Marriott, Separating Mixtures of Normal Distribution, Biometrics, 31, 767-769,1975.

12. GenePix 4000 A User's Guide, Axon Instruments, Inc. 1999.
URL: http://www.axon.com/GN_Genomics.html#software

13. Liu, C. I. Cluster ANOVA with Mixtures(CANOVAM) for Microarray Data.Institute
of Statistics in National Chiao Tung University. The Degree of Master. 2004.

14. Ho, J., Hwang, W.-L., Lu, H. H.-S., and Lee, D. T. Gridding Spot Centers of Smoothly Distorted Microarray Images. IEEE Transactions on Image Processing, to appear. 2005.

15. Li, L., and Lu, H. H.-S. Explore Biological Pathways from Noisy Array Data by Directed Acyclic Boolean Networks. Journal of Computational Biology, 12, 2, 170-185. 2005.

16. Kerr MK, Leiter E, Picard L and Churchill GA. Analysis of a designed microarray experiment. Proceedings of the IEEE-Eurasip Nonlinear Signal and Inage Processing Workshop. June 3-6 2001

17. Kooperberg, C., Fazzio, T. G. , Delrow ,J. J., and Tsukiyama, T., Improved background correction for spotted DNA microarrays," J. Comput. Biol.,vol. 9, pp. 55–66, 2002.

18. McLachlan, G., D. Peel. Finite Mixture Models. Wiley. 2000.

19.O'Neill Paul.Dec. Improved Processing of Microarray Data Using Image Reconstruction Techniques. George D. Magoulas. 2003

20. Speed T. P., Statistical Analysis of Gene Expression Microarray Data, Chapman &Hall/CRC,1 edition , 2003.

21. Schwartz. G. Estimating the dimensions of a model. Ann. Stat.6:461-464 1978.

22. Wu CFJ. On the convergence properties of the EM algorithm. Annals of Statistics 11, 95-103, 1983,.

23. Wu, H.-M., and Lu, H. H.-S. Supervised Motion Segmentation by

Spatial-Frequential Analysis and Dynamic Sliced Inverse Regression. Statistica Sinica, 14, 413-430. 2004.

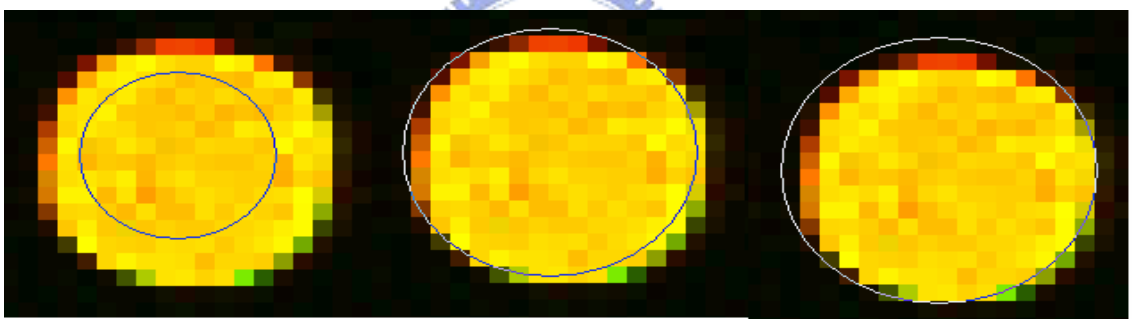
24. Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. Comparison of methods for image analysis on cDNA microarray data. Journal of Computational and Graphical Statistics 11, 108-136, 2002..

25. Yang, Y. H., and Speed, T. P. Design issues for cDNA microarray experiments. Nature Reviews Genetics 3, 579-588. 2002.

26. Yang, Y.H., Buckley, M. J. and Speed T. P. Analysis of cDNA microarray Image. Sep2001.

Appendix I

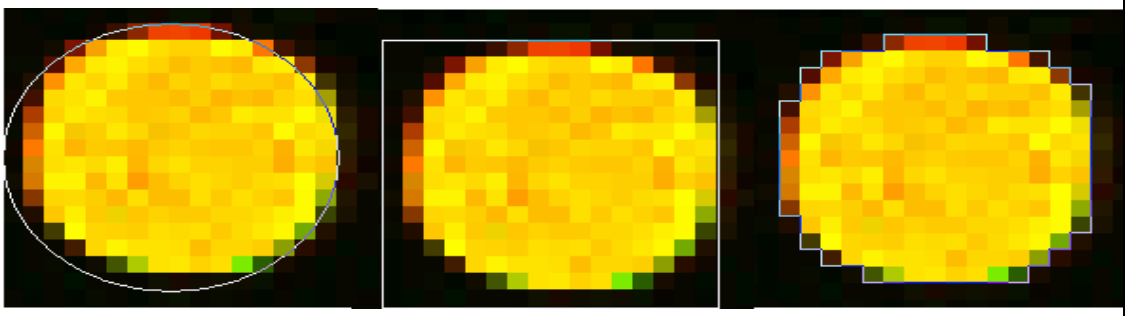
The effects of input arguments in GenePix 6.0 are discussed in this appendix. We will discuss the selections of input arguments for feature diameter and shape of every spot. Three feature diameters of circular boundaries are demonstrated in Figure A.1.1. The feature diameter of 100 μ m is too small to include all foreground pixels in the circular boundary. The circular boundary of feature diameter of 150 μ m will miss some foreground pixels near the boundaries. The circular boundary of feature diameter of 160 μ m will miss few foreground pixels near the boundaries and only few background pixels are included in the segmented region. Hence, the feature diameter of 160 μ m shall be used in this case. The segmentation of circular boundary can be improved by using the segmentation method of irregular boundary..



Feature diameter=100μm	Feature diameter=150μm	Feature diameter=160μm

Figure A.1.1: The segmentation results of three different values of feature diameters for circular boundaries are shown.

Next, we will discuss the effects of segmentation using different boundary shapes. Three segmentation results by the built-in methods of GenePix 6.0 are displayed in Figure A.1.2. Because the shape of spots in one microarray image may not be circular or rectangular, the boundaries of irregular shape can separate the foreground and background more accurately.



Circular Feature	Square Feature	Irregular Feature

Figure A.1.2: The segmentation results of three different boundary shapes are illustrated.

Appendix II

The Gaussian smoothing uses a convolution operator with a Gaussian kernel to remove the noises in signals and images. We consider the following Gaussian distribution function in the two dimensional mask for this study:

$$G(x,y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \qquad \text{(A.II.1)}$$

For example, we can consider the size of mask window = 3 and the standard deviance σ = 1.0. This Gaussian mask of size 3 by 3 is displayed in Table A.II.1. The sum of these 9 mask values could be smaller than 1 and we can divide the original mask value by the sum to obtain the new mask values that are summed to 1. Then, the original 9 intensities will multiple the corresponding mask value and their sum will be used as the new intensity value at the central pixel. This process is applied to all pixels in the image to reduce the effects of noises.

Table A.II.1: A 3 by 3 Gaussian mask is displayed.

$\frac{1}{2\pi}\exp(-\frac{(-1)^2+(-1)^2}{2})$	$\frac{1}{2\pi}\exp(-\frac{(-1)^2+0^2}{2})$	$\frac{1}{2\pi}\exp(-\frac{(-1)^2+1^2}{2})$
$\frac{1}{2\pi}\exp(-\frac{0^2+(-1)^2}{2})$	$\frac{1}{2\pi}\exp(-\frac{0^2+0^2}{2})$	$\frac{1}{2\pi}\exp(-\frac{0^2+1^2}{2})$
$\frac{1}{2\pi}\exp(-\frac{1^2+(-1)^2}{2})$	$\frac{1}{2\pi}\exp(-\frac{1^2+0^2}{2})$	$\frac{1}{2\pi}\exp(-\frac{1^2+1^2}{2})$

Appendix III

The Results of SSRE with GenePix6.0 and NMM in five images are as Table A.III.1 and Table A.III.2.

Table A. III.1: The results of SSRE with GenePix6.0 and NMM for two dye-swap microarrays are listed.

Image		**63T63N**		**63N63T**	
Spike#	T. Ratio	GenePix6.0 Ratio of Median	NMM Ratio of Median	GenePix6.0 Ratio of Mean	NMM Ratio of Mean
Spike1	0.1	0.14	0.14	0.35	0.31
Spike2	0.1	0.05	0.06	0.18	0.16
Spike3	0.2	0.11	0.11	0.28	0.25
Spike4	0.2	0.13	0.13	0.36	0.31
Spike5	0.25	0.23	0.23	0.55	0.48
Spike6	0.25	0.28	0.28	0.71	0.59
Spike7	1	0.94	0.94	2.01	1.74
Spike8	1	1.12	1.02	1.67	1.61
SSRE		**0.5894**	**0.5751**	**14.0938**	**8.8053**

Table A. III.2: The results of SSRE with GenePix6.0 and NMM for two dye-swap microarrays are listed.

Image		62N62T		62T62N	
Spike#	T. Ratio	GenePix6.0 Ratio of Median	NMM Ratio of Median	GenePix6.0 Ratio of Mean	NMM Ratio of Mean
Spike1	0.1	0.11	0.12	0.43	0.41
Spike2	0.1	0.05	0.06	0.22	0.2
Spike3	0.2	0.08	0.09	0.34	0.32
Spike4	0.2	0.10	0.12	0.44	0.41
Spike5	0.25	0.18	0.2	0.66	0.63
Spike6	0.25	0.20	0.23	0.83	0.77
Spike7	1	0.68	0.75	2.36	2.23
Spike8	1	0.48	0.54	2.04	1.97
SSRE		**1.1613**	**0.9149**	**25.2790**	**21.5488**

Table A. III.3: The results of SSRE with GenePix6.0 and NMM for two dye-swap microarrays are listed.

Image		54T54N	
Spike#	T. Ratio	GenePix6.0 Ratio of Median	NMM Ratio of Median
Spike1	0.1	0.32	0.29
Spike2	0.1	0.16	0.14
Spike3	0.2	0.25	0.22
Spike4	0.2	0.32	0.28
Spike5	0.25	0.51	0.45
Spike6	0.25	0.62	0.53
Spike7	1	1.87	1.66
Spike8	1	1.67	1.48
SSRE		**10.0081**	**6.5004**